DOCUMENT RESUME

ED 443 869 TM 031 514

AUTHOR Sireci, Stephen G.; Patelis, Thanos; Rizavi, Saba; Dillingham, Alan M.; Rodriguez, Georgette

TITLE Setting Standards on a Computerized-Adaptive Placement Examination. Laboratory or Psychometric and Evaluative Research Report No. 378.

INSTITUTION Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.

PUB DATE 2000-04-25

NOTE 36p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Adaptive Testing; *College Bound Students; *Computer Assisted Testing; Higher Education; Mathematics; Screening Tests; *Standards; *Student Placement

IDENTIFIERS *Experts; *Standard Setting

ABSTRACT

Setting standards on educational tests is extremely challenging. The psychometric literature is replete with methods and guidelines for setting standards on educational tests; however, little attention has been paid to the process of setting standards on computerized adaptive tests (CATs). This lack of attention is unfortunate because CATs are becoming more widely used, and setting standards on these tests is typically more difficult than setting standards on nonadaptive (linear) tests. This paper discusses some of the issues to be addressed when setting standards on CATs, presents the results of a standard setting study conducted on a computerized adaptive placement test, and discusses the implications of the findings for future research and practice in this area. Thirteen mathematics experts participated in the standard-setting study using ACCUPLACER (College Board) scores. The results of the study suggest that standards can be set on CATs using subsets of items from a CAT item pool, and that methods designed to gather test-centered standard setting data more quickly than traditional methods show promise for setting standards on CATs. (Contains 2 figures, 5 tables, and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 443 869

Setting Standards on a Computerized-Adaptive Placement Examination[1,2]

Stephen G. Sireci[3]

University of Massachusetts at Amherst

Thanos Patelis

The College Board

Saba Rizavi, Alan M. Dillingham, and Georgette Rodriguez

University of Massachusetts at Amherst

S. Sireci

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM031514

[1] Laboratory of Psychometric and Evaluative Research Report No. 378, School of Education, University of Massachusetts, Amherst, MA.

[2] Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April 25, 2000.

[3] The authors thank Kristen Huff and Michael Jodoin for their assistance with some of the literature review and data analysis activities of this research.

BEST COPY AVAILABLE

Abstract

Setting standards on educational tests is extremely challenging. The psychometric literature is replete with methods and guidelines for setting standards on educational tests; however, little attention has been paid to the process of setting standards on computerized-adaptive tests (CATs). This lack of attention is unfortunate because CATs are becoming more widely used and setting standards on these tests is typically more difficult than setting standards on non-adaptive (linear) tests. This paper discusses some of the issues to be addressed when setting standards on CATs, presents the results of a standard setting study conducted on a computerized-adaptive placement test, and discusses the implications of the findings for future research and practice in this area. The results of our study suggest that standards can be set on CATs using subsets of items from a CAT item pool and that methods designed to gather test-centered standard setting data more quickly than traditional methods show promise for setting standards on CATs.

Introduction

Many technically sound and defensible standard setting methods exist in the literature (see Cizek, in press, and Hambleton, 1998 for examples). Many of these methods are designed for specific types of tests, such as tests comprising only multiple-choice items, tests comprising performance tasks, and so forth. However, methods designed for computerized-adaptive tests (CATs) are hard to find. The lack of research on setting standards on CATs is surprising since computerized-adaptive testing is growing in popularity and several large-scale testing programs currently use CATs. Examples of contemporary CATs involving standards include the Registered Nurse exam (Zara, 1997), the Novell Systems Engineer exam (Foster, Olsen, Ford, & Sireci, 1997), and the ACCUPLACER post-secondary placement exams (College Board, 1997).

Although CATs do not require separate standard setting methods, there at least two formidable issues to be addressed by test specialists who set standards on CATs (Sireci & Clauser, in press). First, on a CAT, examinees do not all take the same test. Typically, tests of very dissimilar difficulty are given to examinees of different abilities. Therefore, standards cannot be set on "the" test. Second, CATs typically involve very large pools of items, which may make some of the more popular standard setting methods impractical.

In this paper, we present the results of a standard setting study applied to the Elementary Algebra sub-test of the ACCUPLACER examination program. This exam is a computerized-adaptive test designed to assist colleges and universities in placing students into remedial and beginning college-level math classes. As part of the study, we implemented a modification of the Angoff method (called the item sorting method) that was designed to be easier and quicker for panelists. The purposes of this paper are to: (a) illustrate how the results from a traditional standard setting method can be applied to a CAT, (b) introduce a new standard setting method

designed for CATs, (c) evaluate the new method using several criteria, and (d) discuss the implications of this study for future research and practice in this area.

Method

Description of the ACCUPLACER Elementary Algebra Test

ACCUPLACER is a series of computerized placement exams used throughout the United States for placing students into post-secondary courses. These exams, developed and coordinated by the College Board, were administered to over two million students in 1998. The Elementary Algebra (EA) sub-test was selected for the purposes of this study because it is one of the highest volume tests in the ACCUPLACER battery and some users of this test have expressed concern regarding the validity of their cut-scores.

The ACCUPLACER EA test is most commonly used for determining whether post-secondary students are ready for placement into introductory algebra, intermediate algebra, or college algebra. The EA item bank comprises 120 multiple-choice items measuring three content areas: (a) signed numbers and rationals, (b) algebraic expressions, and (c) equations, inequalities and word problems. All items are calibrated using the three-parameter item response theory (IRT) model (Lord & Novick, 1968). Although there are 120 items in the bank, each student is administered only twelve items. The approximate distributions of items from each content area are between 8 and 17% for content area (a), 42 and 67% for content area (b), and 17 and 50% for content area (c) (College Board, 1997). ACCUPLACER scores are reported on two score scales: percentile score and Total Right Score, which is the score scale on which placement decisions are made. The Total Right Score is the estimated number of items the student is expected to answer correctly if s/he were administered the entire 120-items composing the bank. This score scale (hereafter referred to as the ACCUPLACER score scale) ranges from 0 to 120, with a mean of

48.29, and standard deviation of 26.21 (College Board, 1993). The estimated internal consistency reliability of EA scores is .93 (College Board, 1993).

College Board's Cut-score Recommendations

Given the diversity of students and courses across colleges and universities, the College Board does not mandate specific cut-scores on the ACCUPLACER tests for placing students into specific courses. As described in the ACCUPLACER Coordinator's Guide (College Board, 1997) "it is not possible for the College Board to provide...definitive rules [for schools] to use in...interpretation of scores and placement of students" (p. 43). However, the College Board does provide "typical course placement rules" that are "designed as a guide, a starting point for ... determining placement decisions" (p. 43). For the EA test, the cut-score interval for placing students into "elementary algebra I" is 31—56, the cut-score interval for placing students into "elementary algebra II" is 57—75, and the interval for placing students into "intermediate algebra" is 76—107 (p. 44). Unfortunately, no information is provided regarding how these cut-score intervals were established.

In addition to providing these relatively wide cut-score intervals, the College Board encourages schools to conduct research regarding the most appropriate cut-scores for placement into specific courses. To assist schools in conducting such studies, they also provide a "Placement Validation and Retention Service." Research conducted through the Service focuses on analysis of students' course performance vis-à-vis their ACCUPLACER scores. By analyzing these data, schools can "determine the most appropriate placements for students at a given score level" (College Board, 1997, p. 4).

Although it is true school-specific validation studies are the best way to determine which cut-scores are most appropriate for placement decisions at a particular school, it is also true that

common standards across schools are often desired, particularly for courses considered to be the same across schools. For example, the Commonwealth of Massachusetts recently mandated uniform cut-scores on ACCUPLACER exams for all public post-secondary institutions. In addition, even with the expertise provided by the College Board's Placement Validation and Retention Service, many schools do not have the time or personnel to conduct such studies. Therefore, research into uniform cut-scores on the ACCUPLACER EA test seemed warranted.

Participants

Several criteria were used for selecting mathematics experts to serve as the panelists for the standard setting study. According to Jaeger (1991), "expert judges should have experience with the curricula..., with the knowledge demands imposed by such [courses], and with the capabilities of students who succeed in such [courses]" (p. 4). In selecting participants, we required a minimum of four years mathematics teaching experience at the post-secondary level. We also required teachers to be familiar with (a) the content of different math courses into which students who take the EA test could be placed (i.e., curricular expertise), and (b) the knowledge, skills, and abilities possessed by students who would succeed, and who would not succeed, in such courses. In addition, when selecting panelists, we attempted to achieve representativeness with respect to geographic region, type of school, majority/minority status, and sex.

Thirteen mathematics experts (7 males, 6 females) from two- and four-year colleges across the U.S. were recruited to participate in the study. These panelists came from seven states: California (2), Florida (3), Maryland (2), Massachusetts (1), North Carolina (1), New York (1), Texas (2), and Virginia (1). Although the panel did not reflect perfect U.S. geographic representation, most of the high-volume ACCUPLACER states were represented. These panelists were carefully selected from the network of colleges who were using ACCUPLACER

or had inquired about ACCUPLACER. Nine of the twelve panelists came from schools where ACCUPLACER scores were used for placement decisions. Only one of the panelists never heard of ACCUPLACER before the study. The panelists were well educated and experienced. All had bachelor's degrees, seven had master's degrees, and four had doctoral degrees. Years of post-secondary mathematics teaching experience ranged from five to thirty-four years, with a mean of twenty-one years. Nine of the panelists worked at community colleges, three worked at four-year schools, and one worked at a technical college. Eight panelists reported their primary responsibility was teaching, two reported their primary responsibility was administration, and three reported both. Two panelists were African American; the remaining panelists were Euro-American.

Materials

The ACCUPLACER item bank comprised 120 items; however, six items were retired and two items were accidentally omitted from the standard setting materials. Therefore, the actual number of items reviewed by the panelists was 112. The computer screen shots for these 112 items were printed out on 8.5-by-5.5 inch sheets of paper. The items were partitioned into three subsets of 40, 38, and 34 items, respectively. The partitioning of the items into three subsets made the item review tasks more manageable. In addition, it allowed for an evaluation of the consistency of the standard setting procedures across the three sets of items, and allowed for an evaluation of how well each procedure could work using subsets of items, rather than the entire item bank.

Procedure

The panelists were convened for a two-day meeting. After introductions, the panelists were briefed on the purpose of the study and given an overview of the ACCUPLACER testing

program. They were also given a very brief description of computerized-adaptive testing. Next, the first subset of 40 items (item set A) was distributed and each panelist was asked to answer the items. This task allowed them to get an idea of the difficulty of the items. The panelists took about 25 minutes to answer the 40 items. They were then given the answer key and asked to score their responses.

Discussion of borderline elementary/introductory algebra student

The purpose of the study was to derive two separate cut-scores on the ACCUPLACER score scale. One cut-score was needed for placement into elementary or introductory algebra, the other cut-score was needed for placement into intermediate algebra. The first part of the study focused on the elementary/introductory algebra cut-score. A preliminary description of a student who is "borderline" for placement into elementary/introductory algebra was distributed to the panelists for discussion. The panelists discussed this definition and made several additions, modifications, and deletions. The panelists took about 45 minutes to discuss and reach consensus regarding the description of the student who is borderline for placement into elementary/introductory algebra.

Item sorting task

Next, the panelists were introduced to the item sorting task. This task was designed to be quicker and easier for panelists, relative to the Angoff method. The panelists were instructed to think about the knowledge, skills, and abilities of the borderline elementary/introductory algebra student and review each item in set A. Their task was to sort each item into one of three categories describing the likelihood the borderline student would answer the item correctly. The first category was "very likely to pass this item," the second category was "very likely to fail this item," and the third category was "not sure." The panelists were asked to sort the first five items

in set A. The facilitator then led a discussion of the panelists' sortings. The discussion required people to explain their categorizations of the items. Following the discussion, the panelists were invited to revise their sortings, if they wished. Next, they were asked to sort the remaining items on their own.

Angoff rating task

After the panelists' item sorting data were collected, they were introduced to the Angoff rating task. This introduction included a review of the borderline elementary/introductory algebra student. The panelists were asked to keep this student in mind and then review the first five items in item set A. For each item, they were asked to estimate the probability that this type of borderline student would answer the item correctly. After providing ratings for these five items, their ratings were discussed as a group and panelists were invited to revise their ratings based on the discussions, if they wished. They were then instructed to complete the Angoff ratings for the remaining items in set A on their own.

When all the panelists completed their ratings, they were given a brief description of the classical item difficulty statistic, proportion of examinees answering the item correctly (p-value). They were then given the p-values for all the items in set A and given some time to review these statistics and revise their Angoff ratings, if they desired. After they reviewed these statistics individually, they discussed the Angoff ratings and p-values as a group. This discussion started by asking the panelists which items they wanted to discuss. For the items selected for discussion, each panelist reported their rating and provided their rationale for the rating.

Subsequent steps

After completing the Angoff ratings for item set A, the panelists were told they were to repeat the sorting and Angoff tasks for item sets B and C. They were asked whether they

preferred to do the sorting task or Angoff ratings first. They unanimously agreed to do the sorting task first. Therefore, the steps outlined above were repeated for item sets B and C.

Once the sorting and Angoff data were gathered for all item sets, the panelists reviewed and discussed a draft description of the "student who is 'borderline' for placement into intermediate algebra." The panelists made several changes to this description. Complete consensus was not reached regarding the revised description of the borderline Intermediate algebra student. In particular, some panelists expressed differences of opinion regarding whether these students "are ready to study the concept of a function," or "should understand the concept of a function." There was also some disagreement regarding whether such students should understand the concepts of slope and intercept. About 10 of the 13 panelists were comfortable with the revised description of this borderline student. All panelists were instructed that the goal of the discussion was not to reach consensus, but rather to enable each member to have his/her own understanding of the knowledge, skills, and abilities possessed by this type of student. The panelists asserted they each had a clear understanding of the student who was "just" qualified for placement into intermediate algebra.

Evaluation questionnaire

The panelists also filled out a comprehensive questionnaire to evaluate and compare the different standard setting tasks. This questionnaire asked about the clarity of the descriptions of the borderline students, their degree of confidence in their Angoff and sorting ratings, their perceived difficulty of each task, and their opinion regarding which method (sorting or Angoff) was more effective for setting cut-scores on ACCUPLACER exams. They also rated each aspect of the study (e.g., group discussions, provision of p-values, etc.) with respect to how helpful it was in making their ratings, and indicated whether the amount of time spent on each element of

the study was "not enough," "too much," or "about right."

Data Analysis

Deriving cut-scores from the Angoff data

The Angoff rating data were analyzed in the traditional fashion to derive a cut-score for placement into elementary/introductory algebra, and a cut-score for placement into intermediate algebra. To compute these cut-scores, the item probabilities (i.e., estimated proportion correct for borderline students) were summed for each panelist and the mean sum was computed across panelists, as follows:

$$raw_cutscore = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I} p_{ji}}{J} \qquad [1]$$

where J = the number of panelists (13 in this study), I= the number of items (112 for the full set) and p_{ji} = the Angoff rating for item i provided by panelist j. This procedure provided the cut-scores in terms of the number of items that needed to be answered correctly. To place the number correct cut-score onto the 0-120 ACCUPLACER score scale, the raw cutscore was divided by the total number of items (I), and then multiplied by 120:

$$cutscore = \frac{\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{I} p_{ji}}{J}\right]}{I} X120 \qquad [2]$$

For the complete set of 112 items, I was equal to 112. One of the purposes of the study was to determine what proportion of the item bank was needed for stable estimation of the cut-score. Therefore, separate cut-scores were derived using only items from each of the three sets,

as well as items from pairs of sub-sets. For deriving cut-scores from item sets A, B, and C, *I* was 40, 38, and 34, respectively.

Deriving cut-scores from the sorting data

The sorting procedure was new, and so two different methods were explored for using these data to derive cut-scores. First, the ratings for each item across the 13 panelists were examined. All items with median ratings of "not sure" were identified. The IRT item difficulty parameters (b-parameters) for these items were tabulated and the mean b-parameter was computed. The mean b-parameter for these "not sure" items was taken as the cut-score (on the IRT theta score scale) for the borderline student. Thus, this procedure took advantage of the fact that, in an IRT model, the item difficulty parameters and examinee parameters are on the same score scale.

To transform the cut-scores from the IRT score scale (i.e., theta) to the 0-120 ACCUPLACER score scale, the mean b-parameter for the "not sure" items was entered into the 3PL equation (along with the a-, b-, and c-parameters) to estimate the probability that a borderline student would get each item correct (i.e., P(x=1|θ)):

$$P_i(x=1|\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}} \qquad [3]$$

Thus, the mean b-parameter for the "not sure" items was used as the best estimate of the IRT-score (theta) for the borderline student, and this estimate was used to calculate the probability that the borderline student would answer the item correctly. These probabilities were then summed across all items in the bank to derive the final cut-score, thus:

$$cutscore = \sum_{i=1}^{n} P_i(x=1|\theta) \qquad [4]$$

where n equals the total number of items (112), and θ equals the cut-score on the IRT scale indicated by the mean of the "not sure" items. The cut-scores were divided by 112, then multiplied by 120 to place them on the 0-120 ACCUPLACER score scale.

The second procedure for deriving a cut-score from the sorting data followed the exact same process outlined above, but differed in the way it calculated the theta estimate for the borderline student. Rather than use the mean b-parameter of the "not sure" items, this method performed a contrasting groups analysis using the two other groups of items (i.e., the "very likely to fail" items and "very likely to pass" items). All items with a median rating of "very likely to pass" were coded 1 and the "very likely to fail" items were coded zero. The theta estimate for the borderline student was determined using logistic regression where the dichotomous grouping variable was the criterion, and the vector of IRT b-parameters was the covariate. The logistic regression equation

$$p(x=1) = \frac{1}{1+e^{-(a+bx)}} \qquad [5]$$

where $p(x=1)$ is the probability the item belongs to the "very likely to fail" group, was used to estimate the a and b parameters for the conditional probability function. The theta estimate for the borderline student was derived by setting p=.50, and solving for x. Essentially, this procedure was a contrasting groups analysis where the point on the theta scale that corresponded to the maximum overlap between the b-parameter distributions of "very likely to pass" and "very likely to fail" items was taken as the theta estimate for the borderline student. Livingston and Zeiky (1989, p. 139) introduced this procedure for setting cut-scores when different groups of examinees compose the contrasting groups (see also Sireci, Robin, & Patelis, 1999). In the present study, the procedure was used to contrast groups of items. Given that the examinees'

scores were on the same scale as the b-parameters for these items, the solution for x could be used as the cut-score for students.

Evaluation criteria

Several criteria have been suggested for evaluating standard setting studies (e.g., Cizek, 1996; Hambleton, 1998; Jaeger, 1991; Kane, 1994). In this study, we were interested in both evaluating each method and comparing the methods to one another. The criteria we used in evaluating and comparing the methods were: (a) amount of time it took to gather the data, (b) stability of the cut-scores across sub-sets of items, (c) congruence of the panelists' rating data to the item b-parameters, (d) standard errors of the mean cut-scores across panelists, (e) similarity of the cut-scores across methods, and (f) panelists' evaluation data.

Results

Time to complete ratings

The item sorting task was designed to be quicker than the Angoff task, which would make it more convenient for setting standards on relatively longer tests or tests comprising larger item banks. The amounts of time it took panelists to complete each of the various tasks are presented in Table 1. The time it took to complete each task was computed by subtracting the time all panelists started the task from the time at which the last panelist completed the task. On average, it took the panelists about 35 seconds to answer each item. Including the amount of time it took to train the panelists, discuss their ratings, and revise their ratings, on average, it took the panelists about 38.8 seconds to sort an item and 80.4 seconds to provide an Angoff rating for an item. Thus, the sorting data were collected in less than half the time as the Angoff data (1.00:2.07 ratio). However, one reason it took longer to gather the Angoff data was that the item p-values were given to the panelists during the discussion of the Angoff ratings (since

introducing them during the sorting task would bias the panelists' subsequent Angoff ratings). The panelists' reviews of the item p-values added about 4.1 seconds per item. Subtracting this time from the total yields an average time of about 76.3 seconds to complete an Angoff rating for an item, which is still almost twice as long in comparison to sorting an item (1.00:1.96 ratio).

[Insert Table 1 Here]

As indicated in Table 1, it also took the panelists longer to discuss the typical student who was borderline for placement into an intermediate algebra class than to discuss the typical student who was borderline for placement into introductory/elementary algebra. In fact, the panelists did not reach consensus regarding the description of the borderline intermediate algebra student, due in part to panelists' reports that the content of an intermediate algebra course was not completely consistent from one school to another.

Preliminary Analyses of Item Sorting Data

The sorting task required the panelists to place each item into one of three categories describing the expected performance of the relevant borderline student on that item (i.e., very likely to pass the item, very likely to fail the item, or not sure). After coding the "very likely to pass items" as 1, the "not sure" items as 2, and the "very likely to fail" items as 3, the median rating across the panelists was used to classify each item into one of these three categories. These data are summarized in Table 2 for both types of borderline students. The proportions of items in each category display the expected pattern across the two types of borderline students. The borderline intermediate algebra student was expected to get almost four times as many items correct than the borderline introductory algebra student. Similarly, the borderline introductory algebra student was expected to fail about four times as many items as the borderline intermediate algebra student. These results support the notion that the panelists appropriately

envisioned different types of students when making their ratings for each type of borderline student. It is interesting to note that the number of "not sure" items was almost double for the borderline intermediate algebra student than the borderline introductory algebra student

To determine if the panelists' item sorting data reflected differences among the items in terms of their difficulty, two one-way analyses of variances (ANOVA) were conducted: one for the borderline introductory algebra sorting data, the other for the borderline intermediate algebra data. Each ANOVA used the items' sorting classification as the grouping variable (i.e., 1=very likely to fail, 2=not sure, 3=very likely to pass) and the items' IRT b-parameters as the dependent variable. In both cases, the ANOVAs were statistically significant, indicating the items that were sorted differently also differed with respect to their IRT difficulty parameters. Follow-up tests using the Games-Howell procedure (due to unequal sample sizes and heterogeneous variances) indicated that, in all cases, the mean b-parameters for items in different sorting categories were statistically different from one another. For the borderline introductory algebra data, the proportion of variance in the item difficulties accounted for by the sorting data was .42. For the borderline intermediate algebra data, the proportion of variance accounted for was .37. These analyses are also summarized in Table 2.

[Insert Table 2 Here]

Cut-scores Derived From Item Sorting Method

Cut-scores derived using "not sure" items

Introductory algebra cut-score. As indicated in Table 2, 17 of the 112 items (15.2%) had median ratings of "not sure" after aggregating the item sorting data for the borderline introductory algebra student over the 13 panelists. The b-parameters for these 17 items ranged from –0.88 to 0.50, with a mean of –0.30 and standard deviation of .35. On the 0-120

ACCUPLACER score scale (see equations 3 and 4), the resulting cut-score was 56.97, which fell just outside the upper limit of the College Board's suggested cut-score interval (25—56) for placing students into elementary algebra I (College Board, 1997).

Intermediate algebra cut-score. After aggregating the item sorting data for the borderline intermediate algebra student over the panelists, 31 of the 112 items (27.7%) had median ratings of "not sure." The b-parameters for these 17 items ranged from -.70 to 1.07, with a mean of 0.32 and standard deviation of .46. The resulting cut-score was 78.69, which fell within the College Board's suggested cut-score interval (76—107) for placing students into intermediate algebra (College Board, 1997). It fell just outside the College Board's cut-score interval for elementary algebra II (57—75), which many schools and panelists described as really being an intermediate algebra course.

Cut-scores derived using logistic regression

Introductory algebra cut-score. The introductory algebra cut-score calculated above using the "not sure" items was based on only 17 items. The logistic regression procedure focused on the other two groups of items. The b-parameters for the 16 "very likely to pass" items ranged from –3.83 to .24, with a mean of –1.35 and standard deviation of 1.14. For the 79 "very likely to fail" items, the b-parameters ranged from –1.17 to 1.49 with a mean of .30 and standard deviation of .62. Using equation 5, the theta value that best differentiated these two groups of items was -.91. On the 0-120 ACCUPLACER score scale, the cut-score was 39.95, which is near the midpoint of the College Board's recommended 25—56 cut-score interval for placing students into introductory algebra.

Intermediate algebra cut-score. The logistic regression method for deriving a cut-score from the intermediate algebra sorting data involved 81 items. The b-parameters for the 62 "very

likely to pass" items ranged from –3.83 to 1.11, with a mean of -.48 and standard deviation of .88. For the 19 "very likely to fail" items, the b-parameters ranged from .23 to 1.49 with a mean of .90 and standard deviation of .38. The theta value that best differentiated these two groups of items was .58, which corresponded to a cut-score of 87.96 on the 0-120 ACCUPLACER score scale. This cut-score is below the midpoint of the College Board's recommended 76—107 cut-score interval for placing students into intermediate algebra.

Preliminary Analyses of Angoff Data

As described in the method section, the thirteen panelists independently provided Angoff ratings for the items, discussed these ratings as a group, reviewed the items' p-values, and were given a chance to revise their initial ratings based on the group discussions and item p-values. Both the panelists' initial ratings and revised ratings were collected. On average, the cut-score associated with each panelist did not change much based on the discussions or p-values. For the introductory algebra cut-score, the mean cut-score change from round one to round two was 1.05, with a standard deviation of 2.44. For the intermediate algebra cut-score, the mean change across panelists was 1.96, with a standard deviation of 2.19. The consensus among the panelists, as measured by coefficient alpha, was high for round 1 and increased slightly for round two. For the introductory algebra cut-score, the coefficient alpha of the panelists' ratings was .92 for round 1 and .94 for round 2. For the intermediate cut-score, the alphas were .95 and .97, respectively. The panelists revised (round 2) ratings were used to derive the Angoff cut-scores.

Cut-scores Derived Using Angoff Method

Introductory algebra cut-score. The Angoff cut-scores associated with each panelist ranged from 37.93 to 51.64, with a mean of 43.68 and standard deviation of 4.10. The standard error of the mean was 1.14. The mean cut-score across panelists (43.68) was taken as the

Angoff cut-score for placing students into introductory algebra. This score is near the midpoint of the 25—56 cut-score interval suggested by the College Board for placing students into introductory algebra.

Intermediate algebra cut-score. For the intermediate algebra cut-score, the panelists' cut-scores ranged from 51.27 to 75.87, with a mean of 65.91 and standard deviation of 7.53. The standard error of the mean was 2.09. The mean cut-score was below the lower limit of the cut-score interval suggested by the College Board (76—107) for placing students into introductory algebra. In fact, it was just below the midpoint of the 57—75 cut-score interval recommended for placing students into elementary algebra II.

Comparing the Sorting and Angoff cut-scores

A comparison of the cut-scores derived from the sorting and Angoff data is provided in Table 3. This table presents the cut-scores derived from the sorting data using both the median (not sure) method and the logistic regression method, and the cut-scores derived from the Angoff data. The cut-score intervals recommended by the College Board (1997) are also presented, as are the mean cut-scores used by the 34 to 40 schools who responded to a survey of ACCUPLACER users summarized in Sireci and Dillingham (1999).

For placing students into introductory algebra, the cut-scores derived from the Angoff and logistic regression-sorting methods fall within the cut-score interval suggested by the College Board. The sorting-based median method provided the most stringent cut-score, falling just above the upper limit of the College Board's recommended cut-score interval. The mean cut-scores of the surveyed schools was close to the midpoint of the College Board interval, which is not surprising since the schools probably use these intervals in selecting their cut-scores.

The three different standard setting methods provided very different cut-scores for placing students into intermediate algebra. The Angoff method provided the least rigorous standard (65.91), and the logistic regression method provided the most rigorous standard (87.96). The 22-point difference across these two standards represents almost a full standard deviation on the ACCUPLACER score scale. Looking at the College Board's cut-score intervals, the median and logistic regression cut-scores are consistent with the College Board's recommended cut-score interval for placing students into intermediate algebra. The median-based cut-score is close to the lower limit of this interval, whereas the logistic regression-based cut-score is closer to the midpoint of this interval. The Angoff cut-score is consistent with the interval for placing students into elementary algebra II. It is also the closest to the median of the cut-scores used by the 40 schools who participated in the Texas survey. The variability across methods is consistent with the variability among panelists regarding the description of the borderline intermediate algebra student, and the variability among schools regarding the existence of an elementary algebra II class.

[Insert Table 3 here]

There is no perfect criterion or "gold standard" for determining which cut-scores are "best." Although we do not know what the most appropriate cut-scores should be, it is important to remember that the sorting method was designed to provide similar information to the Angoff method, but in a shorter time frame. Therefore, the large cut-score differences noted across the sorting and Angoff methods are surprising.

Standard errors of the cut-scores

One method for evaluating cut-scores arriving from different standard setting methods is to compare the standard error of the mean (cut-score) across the methods. Computing the

standard error of the mean for the Angoff cutscore is straightforward. Each panelist's cut-score is taken as the unit of analysis and the standard deviation of the individual cut-scores is divided by the square root of the number of panelists. For the median-sorting method, there is no way to compute the standard error of the mean, since cut-scores were not computed separately for each panelist. Separate cut-scores could be computed for each panelist using the logistic regression procedure. This strategy was not chosen initially because we thought it would be better to average the sorting ratings across panelists to classify each item into one of the three categories ("very likely to fail", etc.), than to do a separate analysis for each panelist. One reason for selecting the former option is the possibility that one or more panelists may not use all three categories when classifying the items (but this did not occur in our study). Nevertheless, for the purposes of estimating the standard error of the mean for the logistic regression cut-scores, separate cut-scores were calculated for each panelist. Averaging over the panelists, the cut-scores derived in this fashion were almost identical to the strategy outlined above (i.e., both cut-scores were within .15 points). Therefore, the standard errors associated with these cut-scores are probably appropriate estimates. For the introductory algebra cut-score, the standard error of the mean was 2.14, which was one-point higher than the standard error associated with the Angoff cut-score (1.14). For the intermediate cut-score, the standard error of the mean was 2.83, which was also higher than the standard error associated with the Angoff cut-score (2.09). Therefore, the Angoff cut-scores are more stable with respect to variability across panelists.

In requiring panelists to sort items into one of three categories instead of providing exact probabilities of success on each item, some information is probably lost. The Angoff data represent borderline students' probability of success on an item on a continuous probability scale, whereas the sorting data represent these likelihoods using a discrete, three-category scale.

The lack of consistency in cut-scores between the Angoff and sorting methods could stem from: (a) a loss of information when using the sorting method in place of the Angoff method, or (b) the more time and attention spent reviewing and discussing items when the Angoff method is used. The differences in cut-scores across the median-sorting method and the logistic regression-sorting method could be due to differences in the numbers of items used to estimate the theta score for the borderline candidates. In all cases, the logistic regression method was based on much larger numbers of items (see Table 2).

Relationship to item difficulty

In both the Angoff and sorting tasks, the panelists needed to consider the difficulty of the item and the capabilities of the borderline candidates when providing their ratings. Therefore, there should be a strong relationship between theses ratings and the item difficulty parameters. As a further criterion for evaluating the methods, the average Angoff ratings for each item and the mean sorting rating for each item were compared with the item's b-parameter. These results are summarized in Table 4. Because the panelists had an opportunity to review the item p-values before providing their final Angoff ratings, the correlations between each round of Angoff ratings and the b-parameters were calculated. The round one Angoff ratings were gathered before the panelists reviewed the item p-values; the round two ratings are their final ratings, which were revised based on the group discussions and review of the p-values. The sorting data reflect only changes made based on a group discussion. The panelists' sorting data were not collected before the group discussion, and so the extent to which their initial ratings changed based on these discussions is unknown.

[Insert Table 4 Here]

As would be expected, the correlations for the revised (round two) Angoff ratings are

higher than the correlations observed for round one, which indicates that when panelists revised their ratings, they were influenced by the p-value data. The differences in the percentage of variance accounted for between the round one and round two Angoff ratings were 19.4% and 13.8% for the introductory and intermediate cut-scores, respectively. In comparing the round one Angoff correlations with the sorting correlations, the Angoff data account for more of the variance in the b-parameters. For the introductory algebra cut-score, the round one Angoff data account for about 5% more of the variance; for the intermediate algebra cut-score, the round one Angoff data account for about 12% more variance. These results indicate that the loss of information due to substituting the sorting task for the Angoff task is about 5% to 12%. In comparing the correlations based on the final Angoff ratings with those from the sorting data, the Angoff data account for about 25% more variation of the item difficulties.

Consistency of standards over item subsets

An additional criterion for evaluating the Angoff and sorting methods is the consistency of the cut-scores derived from each method over subsets of items. One of the major purposes of the present study was to discover if stable cut-scores could be derived from subsets of the item pool, in comparison to cut-scores derived from the total item pool. If the cut-scores derived using item subsets were similar to those derived using the entire pool, then future studies could be done using the subsets, which would have obvious economical benefits.

As described in the method section, the 112 items rated by the panelists were split into three subsets, labeled A, B, and C (40, 38, and 34 items, respectively). The division of the item pool into these subsets made coordination of the meeting easier and helped reduce panelists' fatigue. To evaluate the consistency of the cut-scores over subsets, cut-scores were derived using the data from all possible pairs of subsets (i.e., AB, AC, BC) as well as for each individual

subset. The resulting cut-scores are summarized in Table 5, which also presents the maximum deviation of the subset cut-scores from the cut-scores derived using all the items.

When two subsets of items were used (i.e., 72 to 78 items) the resulting cut-scores for introductory algebra were within three-points (about a tenth of a standard deviation) of the cut-score derived using the total set of 112 items. The cut-scores for the logistic regression method were extremely stable in that there was less than a one-half-point difference across the subsets. The maximum difference in cut-scores across the subsets increased noticeably when only a single subset of items was used. These differences ranged from 2.14 (logistic regression) to 5.42 (median-sort). For the intermediate cut-score, the logistic regression method produced the least stable cut-scores over item subsets. The differences in cut-scores across the subsets were over four points, even when two item subsets were used. The Angoff and median-sort cut-scores were more stable. The maximum deviation for the Angoff cut-scores was 2.06 when two subsets were used, and 3.71 when a single subset was used. These values were 1.37 and 3.71 for the median-sort method.

[Insert Table 5 Here]

In general, the cut-scores derived using about 2/3 of the items (i.e., two subsets) were relatively similar to those derived using all of the items. With the exception of the logistic regression cut-scores for intermediate algebra, the maximum cut-score deviation (from the cut-score derived using all items) was about a tenth of a standard deviation. When the cut-scores were derived using only 1/3 of the items, the deviations tended to double, with the largest deviation being about five and one-half points, or two-tenths of a standard deviation. The relationship between number of items used and cut-score stability is summarized in Figures 1 (introductory algebra) and 2 (intermediate algebra). These figures show both the differences in

BEST COPY AVAILABLE

cut-scores across the methods, as well as the stability of the cut-scores for each method.

[Insert Figures 1 and 2 Here]

Panelists' evaluation data

In general, the panelists' survey responses indicated a preference for the sorting task over the Angoff task. On a ten-point rating scale where 1="not at all confident" and 10="completely confident," they expressed slightly higher confidence in the usefulness of the item sorting data for setting standards (median=8) than for the Angoff data (median=7). With respect to task difficulty, using a ten-point rating scale where 1="not at all difficult" and 10="extremely difficult," the item sorting task was rated easier (median=3) than the Angoff task (median=5). However, when asked, "Which method do you think will be more effective for stetting cut-scores on ACCUPLACER?", five panelists responded "Angoff," four responded "sorting," and four responded "not sure." The panelists tended to recommend that both procedures be used in the future. The survey data also indicate the panelists found the group discussions and statistical data helpful in making their ratings, and that adequate amounts of time were allocated for each task. The factors reported by the panelists to be most influential in making their ratings were: their teaching experience, the borderline student descriptions, their perceptions of the difficulty of the items, and the panel discussions.

Discussion

The results of this study are interesting from both research and operational perspectives. From a research perspective, the results provide insight into competing methods for setting standards on CATs. The study also illustrates sound criteria that can be used to evaluate standard setting studies conducted on CATs. From an operational perspective, the results should be informative for users of the ACCUPLACER EA test when evaluating the cut-scores to be

used for placing students into math courses.

Evaluation of Standard Setting Methods

The item sorting task was designed to be quicker and easier for panelists. However, ideally, it should provide cut-scores similar to those garnered from an Angoff study, since it is designed to be a "shortcut" to the Angoff method. The results of this study suggest that the item sorting method cuts the time required to rate items in half, relative to the "traditional" Angoff method. Thus, it is appealing in that more items can be evaluated in a given time period. Shortening the time to gather standard setting data is desirable in a CAT environment when the number of items composing the item bank is large. However, the finding that the item sorting method produced some cut-scores that were noticeably different than the Angoff method is troubling, and precludes us from endorsing it at this time.

The logistic regression sorting cut-score and the Angoff cut-scores were similar for the elementary algebra standard, which suggests that the method may have some promise and argues for further research on the item sorting method. However, for the intermediate algebra standard, the cut-scores across the three methods were very different. Given the standard deviation of the panelists' cut-scores, the amount of time they took to discuss the borderline intermediate algebra student, and their lack of consensus in describing this student, it is clear that deriving a uniform cut-score for the intermediate algebra placement decision is difficult. It is interesting to note that the College Board provided three recommended cut-score intervals: one for introductory algebra I, another for introductory algebra II, and a third for intermediate algebra (see Table 3). It may be that some panelists envisioned their intermediate algebra course to be a lower level course (such as introductory algebra II) and others considered it to be a relatively higher-level course (post introductory algebra II). The Angoff method probably averages out such differences in

perception better than the sorting methods.

The panelists' survey responses indicated a slight preference for the sorting task. In addition, the sorting data were gathered in about half the time as the Angoff data. However, the sorting data exhibited lower correlations with the item difficulty data, which could indicate too much information is lost when substituting this procedure for the Angoff method. This loss of information could be due to the use of a discrete three-point scale in place of the continuous probability scale, or it could be due to spending less time in reviewing, discussing, and rating the items.

One way to evaluate the relative merits of the specific cut-scores arising from the different methods is to compare them to the results of the validity studies conducted by specific schools. If consistency in cut-scores derived from these studies is found across schools, then these "common" cut-scores could be used to evaluate the cut-scores resulting from the different methods investigated in this study. Unfortunately, the research database for the validity studies is not fully coordinated, and the soundness of these various school-specific validity studies is unknown. In the absence of these data, there are two observations that suggest the Angoff cut-scores are preferable to those derived from the sorting data. First, the Angoff cut-scores had smaller standard errors relative to the cut-scores derived using logistic regression. Second, higher correlations were observed between the (round one) Angoff ratings and the b-parameters relative to the sorting ratings. Nevertheless, future studies should strive to gather external criterion data, such as students' course grades, to help evaluate which cut-scores are best.

Proportion Of Items Needed To Derive Cut-scores

One encouraging finding from this study is the relative stability of the cut-scores across sub-sets of items. When only about two-thirds of the items were used to derive the cut-scores,

they were all within three points (just over a tenth of a standard deviation) of one another. In some cases, the cut-scores derived using two-thirds of the items were within one point of the cut-scores derived using all of the items. For the introductory algebra cut-score, the logistic regression method exhibited the greatest cut-score consistency over sub-samples of items. However, it exhibited the worst consistency for the intermediate cut-score. Further research is needed to determine if the logistic regression approach is better than the simpler "median" approach for analyzing item sorting data.

The cut-scores derived from only a third of the items exhibited greater deviation from the cut-scores derived from all the items, but even then, the deviations were relatively small (about two-tenths of a standard deviation). These results support the idea of using subsets of items, rather than the entire item pool for deriving cut-scores. Clearly, the use of item subsets for estimating cut-scores is promising, and deserves further study.

Implications For ACCUPLACER EA Cut-scores

In the absence of a valid criterion for evaluating cut-scores derived from the Angoff, median-sort, and logistic regression-sort methods, the cut-scores based on the Angoff method seem most defensible. The panelists spent more time discussing the Angoff ratings, and these ratings reflect their revisions based on the "reality check" provided by a review of the item p-values. For the introductory algebra cut-score, the cut-scores derived from all three methods fell near or within the cut-score interval suggested by the College Board (1997). This finding provides evidence in support of this interval. The College Board could also provide a "point estimate" of 44 (the Angoff cut-score rounded up) to schools that are not sure where to set their cut-point within this broad interval. Given the panelists' consensus of the knowledge, skills, and capabilities possessed by a student "just ready" for placement into introductory algebra, the

Angoff cut-score of 44 represents the average cut-score generated from a carefully selected national panel of math experts.

The Angoff cut-score for placing students into intermediate algebra fell within the College Board's recommended cut-score interval for introductory algebra II. This result does not support the College Board's current cut-score intervals for placing students into introductory algebra II or intermediate algebra. Further clarity is required regarding the differences between introductory algebra II and intermediate algebra. Before recommending a cut-score for placement into intermediate algebra, research should be conducted into the methods used to arrive at the College Board's current recommended cut-score intervals. The procedures used to derive the current cut-score intervals should be weighed against the information provided in the current study to decide if adjustments to the current cut-score intervals are needed and if a point estimate should be provided.

Limitations of the Study

From a research design perspective, this study was limited in that there was no condition that involved the item sorting task coupled with a review of the item p-values. In addition, the item sorting task preceded the Angoff task, which could have introduced a carry-over effect. Thus, the present study did not represent a totally fair comparison of the Angoff and item sorting methods. Future research should try to counterbalance the item p-value review step, and order in which the tasks are conducted, across the two methods.

The present study was also limited in that only a single panel of experts was used and the methods were evaluated using only a single test. Future studies should focus on other types of tests and use independent panels of experts, if possible. The math experts in the current study had no problem understanding the probability scale underlying the Angoff ratings. These math

experts were also very quick in providing their ratings. Content experts from other fields may have more difficulty understanding the Angoff task.

Conclusions

In our view, the amount of time discussing the item difficulties and the borderline candidates were the most important factors influencing panelists' Angoff and item sorting judgments. The sorting method is quicker than the Angoff method, but the results of this study do not support its use in lieu of the Angoff method. However, given the limitations of this study, future research on the item sorting method is needed. If valid cut-scores can be produced more quickly than the Angoff method, this procedure could be useful when the number of items composing a CAT item pool is large.

The results of this study suggest standards can be set on CATs using subsets of items rather than the entire item pool. This finding is encouraging, given that more and more educational tests are becoming computerized-adaptive. However, more research is also needed to determine the number of items necessary for reliable estimation of cut-scores.

The ACCUPLACER tests are designed to serve a wide constituency. The results of this study can be used to provide more information to this constituency for using ACCUPLACER scores for making math placement decisions. Future studies are recommended for the other tests in the ACCUPLACER battery.

References

Cizek, G., (Ed.) (in press). Standard setting: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (1996). Standard-setting guidelines. Educational Measurement: Issues and Practice, 15 (1), 13–21, 12.

College Board (1993). ACCUPLACER: Computerized placement tests: Technical data supplement. New York, NY: Author.

College Board (1997, January). ACCUPLACER Program Overview: Coordinator's Guide. New York, NY: Author.

Foster, D., Olsen, J., Ford, J., & Sireci, S.G. (1997, March). Administering computerized certification exams in multiple languages: Lessons learned from the international marketplace. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Hambleton, R. K., (1998). Setting Performance Standards on achievement tests: meeting the requirements of Title I. In L.Hansche (ed.), Handbook of standard setting. Washington, DC: Council of State School Officers.

Jaeger, R. M. (1991). Selection of judges for standard setting. Educational Measurement: Issues and Practice, 10 (2), 3–6, 10, 14.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. Applied Measurement in Education, 2 (2), 121–141.

Lord, F, M, & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison Wesley.

Sireci, S. G., & Clauser, B. E. (in press). Issues to be considered in setting standards on computerized-adaptive tests. In G. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum.

Sireci, S. G., & Dillingham, A. (1999). ACCUPLACER Cut-off score summary report. Unpublished report, School of Education, University of Massachusetts, Amherst, MA.

Sireci, S. G. Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. Applied Measurement in Education, 12 (3), 301–325.

Zara, A. R. (1997, March). Administering and scoring the computerized adaptive test. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Table 1

Duration of Standard Setting Tasks

Task	Duration (in minutes)
Answer test items (112 items)	65
Discussion of "Borderline Elementary/Introductory Algebra Student"	45
Discussion of "Borderline Intermediate Algebra Student"	70
Collect item sorting data for Elementary/Introductory cut-score	78
Collect Angoff data for Elementary/Introductory cut-score	180
Collect item sorting data for Intermediate cut-score	75
Collect Angoff data for Intermediate cut-score	120
Total time for all sorting ratings	145
Total time for all Angoff ratings	300
Complete evaluation questionnaire	30

Table 2

Summary of ANOVA Results on Sorting Data

(a) Borderline Introductory Algebra Student

	Very likely to pass	Not sure	Very likely to fail
# of items	16	17	79
Mean	-1.35	-.30	.30
St. deviation	1.14	.35	.62

$F_{(2,109)}=39.92$, $p<.001$, $\eta^2=.42$

(b) Borderline Intermediate Algebra Student

	Very likely to pass	Not sure	Very likely to fail
# of items	62	31	19
Mean	-.48	.32	.90
St. deviation	.88	.46	.36

$F_{(2,109)}=32.16$, $p<.001$, $\eta^2=.37$

Table 3

Comparison of Cut-Scores

Method	Introductory Algebra Cut	Intermediate Algebra Cut
Angoff	43.68	65.91
Sorting (Logistic Regression)	39.95	87.96
Sorting (Median)	56.97	78.69
College Board's Recommended Cut-score Interval	25—56 (Elementary Algebra I)	57—75 (Elementary II) 76—107 (Intermediate)
Mean School Survey Cut-score[a]	39.7	59.68

[a]Summarized in Sireci and Dillingham (1999). Sample sizes were 34 and 40 schools for the introductory and intermediate algebra cut-scores, respectively.

Table 4

Correlations Among Panelists' Data and IRT b-parameters

	Sorting	Angoff Round 1	Angoff Round 2	r^2 difference Ang1-Sort	r^2 difference Ang2-Ang1	r^2 difference Ang2-Sort
r Introductory	.666	-.700	-.827			
r Intermediate	.687	-.768	-.853			
r^2 Introductory	.444	.490	.684	.046	.194	.240
r^2 intermediate	.472	.590	.728	.118	.138	.256

Table 5
Cut-score Consistency Over Item Subsets

		Introductory Algebra Cut			Intermediate Algebra Cut		
Item Subsets	Number of Items	Angoff	Sorting (LR)	Sorting (Median)	Angoff	Sorting (LR)	Sorting (Median)
ABC	112	43.68	39.95	56.97	65.91	87.96	78.69
AB	78	41.53	39.93	59.94	64.98	91.11	78.18
AC	74	43.53	39.89	55.36	64.88	87.82	80.06
BC	72	46.17	40.42	56.22	67.97	83.38	77.37
A	40	39.21	42.09	58.34	62.20	92.76	80.79
B	38	43.98	40.89	62.39	67.91	87.42	74.98
C	34	48.62	39.97	52.91	68.04	-----	79.33
Maximum Absolute Subset Deviation							
2 Subsets		2.49	0.47	2.97	2.06	4.58	1.37
1 Subset		4.94	2.14	5.42	3.71	4.80	3.71

Figure 1

Item Subset Cut-Score Consistency: Introductory Algebra

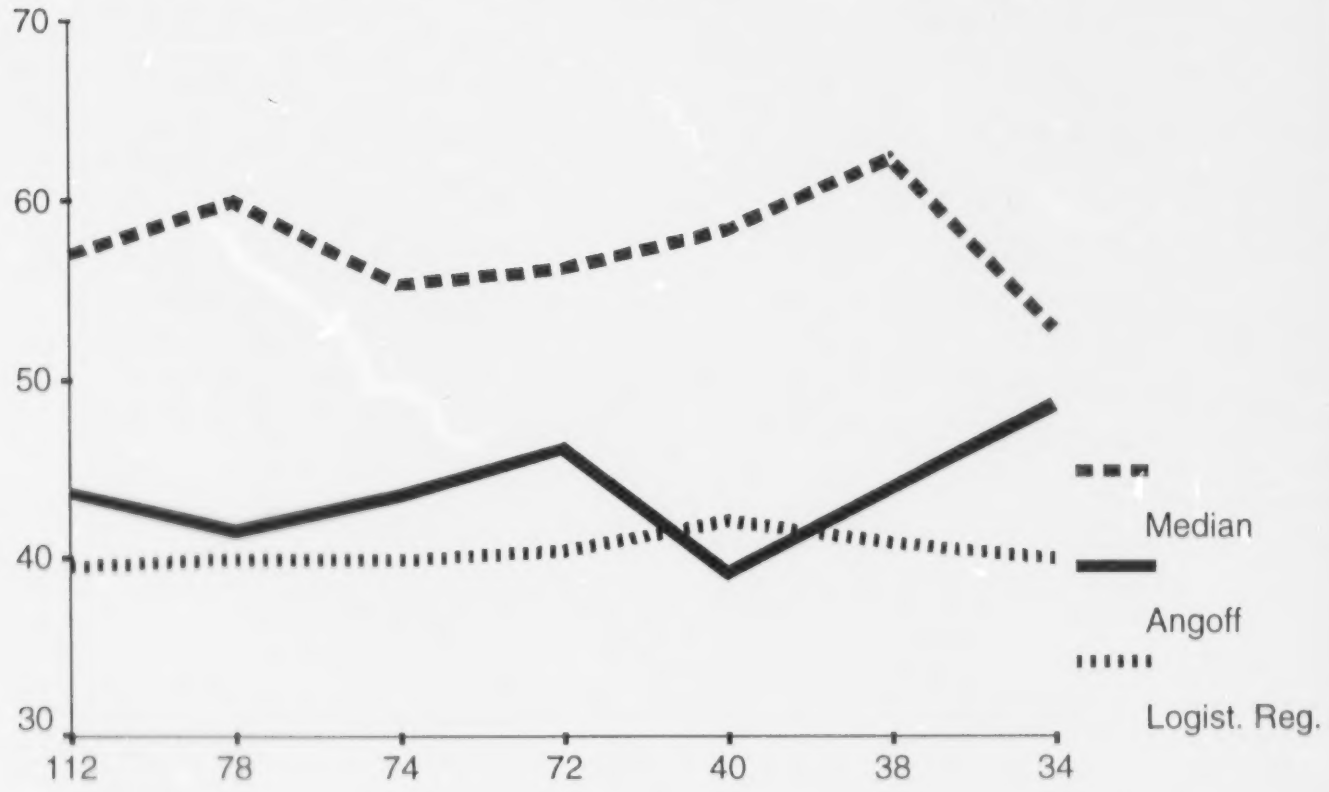

Figure 2

Item Subset Cut-Score Consistency: Intermediate Algebra